Review

ibc Interdisciplinary Bio Central

Open Access, Open Review Journal

Directional adjacency-score function for protein fold recognition

Muyoung Heo[1, 3], Mookyung Cheon[1], Suhkmann Kim[2], Kwanghoon Chung[1, 4], and Iksoo Chang[1,*]

[1]Creative Research Initiatives Center for Proteome Biophysics, Department of Physics, Pusan National University, Pusan 609-735, Korea

[2]Department of Chemistry, Pusan National University, Pusan 609-735, Korea

[3]Current address: Department of Chemistry and Chemical Biology, Harvard University, Boston, MA 02138, USA

[4]Current address: Ministry of Education, Science, and Technology, Gwacheon 427-715, Korea

Subject areas: Bioinformatics, Protein folding

Author contribution: I.C. designed the research, M.H., M.C., S.K., K.C. performed the calculation and analyzed the result, and I.C. wrote the paper.

***Correspondence** and requests for materials should be addressed to I.S.C. (chang@random.phys.pusan.ac.kr).

Editor: Keun Woo Lee, Gyeongsang National University, Republic of Korea

Received June 09, 2009;
Accepted June 23, 2009;
Published June 25, 2009

Citation: Heo, M., et al. Directional adjacency-score function for protein fold recognition. IBC 2009, 1(2):8, 1-6. doi:10.4051/ibc.2009.2.0008

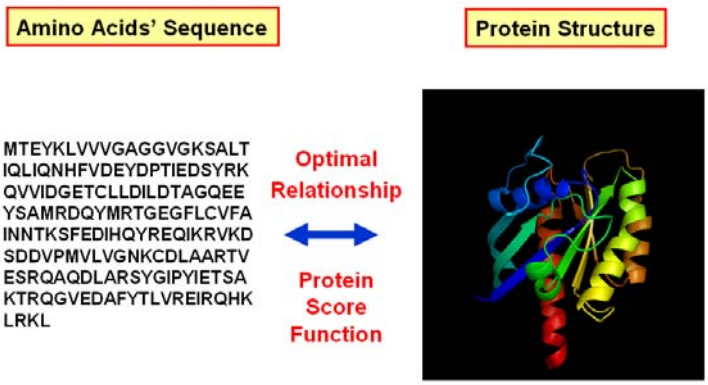
Funding: National Creative Research Initiatives program (Center for Proteome Biophysics) of the ministry of education, science, and technology/Korea Science and Engineering Foundation, Korea.

Competing interest: All authors declare no financial or personal conflict that could inappropriately bias their experiments or writing.

SYNOPSIS

***Introduction*:** It is a challenge to design a protein score function which stabilizes the native structures of many proteins simultaneously. The coarse-grained description of proteins to construct the pairwise-contact score function usually ignores the backbone directionality of protein structures. We propose a new two-body score function which stabilizes all native states of 1,006 proteins simultaneously. This two-body score function differs from the usual pairwise-contact functions in that it considers two adjacent amino acids at two ends of each peptide bond with the backbone directionality from the N-terminal to the C-terminal. The score is a corresponding propensity for a directional alignment of two adjacent amino acids with their local environments.

***Results and Discussion*:** We show that the construction of a directional adjacency-score function was achieved using 1,006 training proteins with the sequence homology less than 30%, which include all representatives of different protein classes. After parameterizing the local environments of amino acids into 9 categories depending on three secondary structures and three kinds of hydrophobicity of amino acids, the 32,400 adjacency-scores of amino acids could be determined by the perceptron learning and the protein threading. These could stabilize simultaneously all native folds of 1,006 training proteins. When these parameters are tested on the new distinct 382 proteins with the sequence homology less than 90%, 371 (97.1%) proteins could recognize their native folds. We also showed using these parameters that the retro sequence of the SH3 domain, the B domain of Staphylococcal protein A, and the B1 domain of Streptococcal protein G could not be stabilized to fold, which agrees with the experimental evidence.



Keywords: protein score function, stabilization of native states, perceptron learning, protein threading, retro protein

Introduction

Protein energy (score) function which can stabilize the native folds of proteins is an important ingredient to study the protein folding problem. The nature of real interaction energies between atoms in a protein is complicated, thus one of the simple and efficient way to study the protein folding problem is to employ a coarse-grained description of amino acids in a sequence after integrating out the details of a protein. Each amino acid is considered as an isotropic point sphere centered at C_α position along the backbone of a protein, and the protein energy function is computed by adding the interaction energies between amino acids. It is widely accepted that the amino acids sequence possesses the essential feature of a protein and that its native structure corresponds to that of minimum free energy (Anfinsen 1973; Maiorov and Crippen 1992; Wolynes et al. 1995; Fersht 1998; Baker 2000; Mayor et al. 2003). Therefore it is important to develop a protein energy function which depends on the sequence of amino acids and their properties. If it is successful, such a function should be able to stabilize the native folds of as many proteins as possible. The basic idea is to develop a score function which can assess the compatibility of amino acids sequence to the structures so that the native state of a protein can obtain the lowest score against when the same sequence of the native state is housed in the competing structures (Friedrichs and Wolynes 1989; Bowie et al. 1991; Goldstein et al. 1992; Vendruscolo et al. 1999; Dima et al. 2000; Salvi and DeLosRios 2003).

Significant achievements were made to construct protein energy functions, based on the known structures of proteins in the Protein Data Bank (PDB) (http://www.rcsb.org/pdb) (Hobohm et al. 1992; Hobohm and Sander 1994; Holm and Sander 1996), which brought us a good understanding of the properties of proteins. Bowie et al., (1991) developed the one-body score function which can probe the compatibility of a sequence against structures of proteins. This has stimulated many fold recognition methods to have better assess for the native folds (Godzik et al. 1992; Jones et al. 1992; Sippl and Weitckus 1992; Bryant and Lawrence 1993; Ouzounis et al. 1993; Wilmanns and Eisenberg 1993; Fischer et al. 1996). Miyazawa and Jernigan (MJ) (1985, 1996) constructed 20×20 matrix for the pairwise contacts of 20 amino acids using the quasi-chemical approximation. The pairwise contact is established when the distance between two amino acids is within a threshold distance (for example 6.5Å) in 3D-space, but the chemical distance between them along the backbone of a protein is larger than three peptide bonds unit. MJ matrix is symmetric and the number of independent parameter is 210. In MJ construction one basically counts the frequency of pairwise contacts of two amino acids which are within the threshold distance in the protein structure. Once a set of proteins used in the statistical counting of pairwise contacts is given, the values of 210 independent parameters are fixed. When MJ matrix is subject to stabilize the native states of the same set of proteins, it could not stabilize all proteins simultaneously but stabilize 70~80% of them. It, however, explained energetic characteristics of proteins and has given a great impact on solving the protein folding problem. Zhang and Kim (2000) proposed a similar approach by expanding the parameter space of amino acids after considering the secondary structure (α-helix, β-sheet, other) of a protein. Each amino acid can be in one of three secondary structures, thus the pairwise contact matrix becomes 60×60 matrix and the number of independent parameters increases to 1,830. Their contact matrix could stabilize more than 97% of the 316 testing proteins, but not all of them.

The aim of a construction of protein energy function is to come up with the appropriate global energy function which can stabilize all the native states of as many proteins as possible so that the sequence in the native structure acquires the lower energy than in the decoy structures. In order to achieve the stabilization of the native states of many proteins, the optimization schemes such as the Z-score method (Friedrichs and Wolynes 1989; Goldstein et al. 1992) and the perceptron learning method (Krauth and Mezard 1987; Vendruscolo et al. 1999; Dima et al. 2000; Salvi and DeLosRios 2003) have been employed without a complete success. Recently Vendruscolo et al. (1999) and Salvi and DeLosRios (2003) tried to determine the pairwise contact energy parameters (20×20 matrix) employing the perceptron learning of training proteins. They showed that only a small set of proteins could be stabilized, and sometimes could not stabilize even a single protein. They concluded that it is not possible to construct a global protein energy function based on only the pairwise contacts of amino acids. The parameterization of amino acids used in these works was only the identity of 20 amino acids, and it was too simple to account for the energetics of proteins. However, the inclusion of the local environments of amino acids, such as the local secondary structure, the solvent accessibility (hydrophobicity), and the polarity, would improve the stabilization capability of the proposed energy function for many proteins. In fact, it was shown that one could construct one-body energy function in terms of the propensities for amino acids to be at the given local environments, which stabilized simultaneously all native states of 600 proteins using the perceptron learning and the protein threading (Chang et al. 2001). One could also construct the two-body energy function after taking into account of the local environments of amino acids and postulate the various forms of energy functions. Whether the postulated forms of energy functions are amenable to stabilize all the native states of training proteins in the perceptron learning process is a tough measure for the success of design and construction of a global protein energy function. One of the simple way to construct such a function is to include the information of local environments of amino acids into the pairwise contact function. After parameterizing the local environments of amino acids into 9 categories depending on three secondary structures and three kinds of hydrophobicity (solvation) of amino acids, the 16,290 independent parameters of pairwise contact (180×180) matrix were determined by the perceptron learning and the protein threading. These could stabilize all native states of 1,006 proteins with 30% homology (Heo et al., 2004).

In this paper we propose a new two-body score function which can also stabilize all native states of 1,006 proteins simultaneously. This two-body score function differs from the usual pairwise contact functions of Miyazawa and Jernigan (1985, 1996), Vendruscolo et al. (1999), Dima et al. (2000), and Salvi and DeLosRios (2003) in that it considers two amino acids which are adjacent each other at two ends of each peptide bond with the backbone directionality from the N-terminal to the C-terminal. The score is a corresponding propensity for such a directional alignment of two adjacent amino acids with their local environments. We show that the construction of a directional adjacency-score function was achieved using 1,006 training proteins with the sequence homology less than 30%, which include all the representatives of different protein classes. After parameterizing the local environments of amino acids into 9 categories depending on three secondary structures and three kinds of hydrophobicity of amino acids, the 32,400 adjacency-scores of amino acids could be determined by the perceptron learning and protein threading. These could stabilize all native folds of 1,006 training proteins simultaneously. When these parameters are tested on the new distinct 382 proteins with the sequence homology less than 90%, 371 (97.1%) proteins could recognize their native folds. Using these adjacency-score parameters, we showed also that the retro sequence of the SH3 domain, the B domain of Staphylococcal protein A and the B1 domain of Streptococcal protein G can not be stabilized to fold, which agrees with the experimental evidence (Lacroix et al. 1998).

Results and Discussion

Directional adjacency-score function

Given a sequence of amino acids, we need to have the directional adjacency-score function which can access the fitness of a sequence to the native structure or decoy structures. We employ the coarse-grained representation of amino acids and their local environmental information. One may construct the simple adjacency-score function in terms of the propensities of two amino acids being adjacent at two ends of each peptide bond along the directional backbone from the N-terminal to C-terminal. The basic strategy is to determine these propensities such that the adjacency-score of a sequence in the native structure is always lower than in the competing decoy structures. This criterion should also apply to the set of many proteins simultaneously. The directional adjacency-score function we propose is the following:

$$H_A(s,\Gamma) = \sum_{i,j}\sum_{k,l} n_A(i,j;k,l)\varepsilon_A(i,j;k,l), \qquad (1)$$

where ***H*** is the directional adjacency-score function which is a measure how good a sequence *s* is housed into the structure **Γ**. The elements in the sum are the environment dependent two-body adjacency-scores for two amino acids residing at two ends of each peptide bond where $n_A(i,j;k,l)$ is the number of pairs of two adjacing amino acids of types *i, j* found in the local environment *k, l* respectively and $\varepsilon_A(i,j;k,l)$ is the propensity associated with it. Here, we considered a direction of a backbone in a protein such that N-terminal is a starting point and C-terminal is an ending point. In our convention the amino acid of type *i* is located at the left end (toward the N-terminal) and type *j* at the right end (towards the C-terminal) of a peptide bond. Therefore, we take into account of the non-symmetric nature of the alignment of amino acids sequence. Once the structures of proteins are given, $n_A(i,j;k,l)$ are determined from PDBs. Our aim is to extract the directional adjacency-score parameters in the 180×180 matrix $\varepsilon_A(i,j;k,l)$ to ensure the simultaneous stabilization of the native folds with respect to a set of decoy structures. Since this matrix is non-symmetric, the number of independent parameters is 32,400.

Local environments of amino acids and a training set of proteins

We classified the local environments of amino acids in the protein structure into 9 categories: Each amino acid can be found in one of three secondary structures (α-helix, β-sheet, and other). The solvent exposed ratio of amino acid is calculated using Richards' algorithm (Lee and Richards 1971; Pattabiraman 1995) as the ratio between the solvent accessible area of each amino acid, X, in its native structure and corresponding area in Gly-X-Gly extended structure. The values of the solvent exposed ratios < 10%, 10-50%, and > 50%, capturing the degree of a hydrophobicity (solvation), were classified into three classes of small, medium, and large exposure respectively. Once this environmental classification of amino acids is done, 3D structural information of a sequence is transformed into 1D string of local environmental parameters. Each protein conformation, namely a sequence housed in a given structure, is now represented by a string of {(*i,m*)}. Therefore the directional adjacency-score function Eq.(1) provides a quantitative measure of the propensity for two amino acids being adjacent at two ends of each peptide bond along the directional backbone of a protein with their corresponding local environments.

We used a training set of 1,006 proteins from the PDB select (http://www.cmbi.kun.nl/gv/pdbsel) and WhatIf (http://www.cmbi.kun.nl/whatif) (Hobohm et al. 1992; Hobohm and Sander 1994; Holm and Sander 1996). In fact, there are 3032 representative proteins with 30% sequence homology in PDB select and WhatIf, covering all different classes according to the Structural Classification of Proteins (SCOP) classification, which were selected from all the known structures of proteins by the all-against-all Smith/Waterman alignment between chains. Among these proteins we selected those (1) whose structure were obtained by x-ray crystallography, (2) which do not have the non-standard amino acids, (3) which are not the disconnected chains, and (4) which are not the structures of mutant. As a result, we have a training set of the non-redundant 1,006 proteins, whose length ranges from 53 to 994 amino acids. The same criterion is used for selecting 382 test proteins with 90% sequence homology which are distinct from 1,006 training proteins and will be used for a stringent threading test of the learned score parameters.

Perceptron learning of the directional adjacency-score parameters

We first generate the decoys of each protein by the gapless threading of 1,006 training proteins on themselves. The sequence of each protein is threaded on the structure (environments) **Γ** of all proteins, with the equal or the longer length than a target protein, out of 1,006 proteins. The solvent accessible area of amino acids mounted on a threaded fragment was approximated to be the same as that in the longer protein from which the fragment was taken. The total number of decoys for 1,006 training proteins is about 78.2 million, and each decoy has to satisfy the following inequality to stabilize the native structures of all 1,006 training proteins (Krauth and Mezard 1987; Chang et al. 2001):

$$\sum_{i,j=1}^{20}\sum_{k,l=1}^{9}[n_A(i,j;k,l)^D - n_A(i,j;k,l)]\varepsilon_A(i,j;k,l) > 0, \qquad (2)$$

where $n_A(i,j;k,l)^D$ and $n_A(i,j;k,l)$ are the occurrence of adjacent pair (*i, j; k, l*) in the decoy *D*(=1,2,…,78.2 million) and in its native structure, respectively. Our aim is to determine and optimize 32,400 parameters of $\varepsilon_A(i,j;k,l)$ to ensure that 1,006 training proteins with the known native structure have the lower scores than when their sequences are housed in the decoy structures.

The basic ingredient to determine the optimal $\varepsilon_A(i,j;k,l)$, instead of solving all 78.2 million inequalities, is the following. For each training protein there are many decoys generated from a protein threading. We impose the condition that native score of a given protein must be lower than (1) the average score of a random sequence on its own native structure with the same composition of amino acid (Seno et al. 1996; Micheletti et al. 1998) and (2) the average score of a sequence on the decoy structures. The former generates 1,006 inequalities and 1,005 inequalities from the later. We first solve these 2,011 inequalities by the perceptron learning, whose solution guides the approximate direction of the ultimate solution $\varepsilon_A(i,j;k,l)$ in 32,400-dimensional parameter space. Using these learned $\varepsilon_A(i,j;k,l)$, we perform a threading test for comparing the scores of all 78.2 million decoys with their native state score. The number of failed decoys whose scores are lower than their native state score is 220 out of 78.2 million. The inequalities from the failed decoys are added to the previous 2,011 inequalities for all of which the perceptron learning, taken the (learned) $\varepsilon_A(i,j;k,l)$ as the initial condition, is performed again to find the new solution for $\varepsilon_A(i,j;k,l)$. We tried to achieve the maximum stability of native state against the competing decoys by maximizing the gap between the native scores of 1,006 training proteins and their failed decoys. Now the second threading test of 1,006 training proteins with the new $\varepsilon_A(i,j;k,l)$ produces the new set of failed decoys adding to the previous set of inequalities. We iterated the procedure of (i) perceptron learning for updating score parameters, (ii) protein threading to add new inequalities until the number of failed decoy to add becomes zero. When this is achieved, the total number of inequalities to solve is

TABLE I. The list of 11 failed proteins out of 382 test proteins.

PDB	N_{AA}	N_{td}	N_{fd}	PDB	N_{AA}	N_{td}	N_{fd}
1IG7	58	78,676	5	1HRO	105	61,461	26
1FYN	62	77,160	21	1CO6	107	60,764	1
1PGX	70	74,154	1	1HE7	107	60,764	22
1MHO	88	67,522	6	1G96	111	59,402	15
1BWO	90	66,795	7	1RAV	124	55,174	1
1HPO	99	63,567	2				

The number of failed decoys (N_{fd}) is within the lowest 0.1% of the total number of decoys (N_{td}) for each protein showing that the native folds are almost stabilized even for the failed proteins. N_{AA} is the length of each protein.

2,235. Although the solution $\varepsilon_A(i, j; k, l)$ from solving 2,235 inequalities satisfies all 78.2 million inequalities, it is neither unique nor optimized. The optimization strategy is to push the scores of competing decoys as further away as possible from the native state score so that the maximum stabilities of the native states of 1,006 training proteins are achieved. For this purpose we identify the competing decoys (among all 78.2 million decoys) whose score gap from their native state score is smaller than the minimum gap of 2,235 decoys. Again we add the inequalities for these competing decoys to the previous 2,235 inequalities, and learn the optimized solution $\varepsilon_A(i, j; k, l)$. We also iterate the procedure of perceptron learning and protein threading until the number of competing decoys (among all 78.2 million decoys) whose gap is smaller than the minimum gap of previous inequalities becomes zero, which resulted in just solving 4,013 inequalities. We could optimize 32,400 adjacency-score parameters simultaneously which stabilize 100% of the native states of 1,006 training proteins.

Threading test of the directional adjacency-score parameters and stabilization capability for the new distinct proteins

After we succeeded in learning the directional adjacency score parameters, we check the stabilization capability of our parameters for the native folds of the new distinct proteins. The threading test of 382 new distinct proteins on themselves using our learned score parameters showed that the native folds of 371 (97.1%) proteins could be stabilized, and there are only 107 failed decoys out of the total 12.1 million decoys. TABLE I lists 11 failed proteins, and the number of failed decoys for each of them is within the lowest 0.1% of the total number of decoys. In view of the fact that we chose our 382 test proteins which are 90% homologous in order to perform a stringent threading test, the success ratio of more than 97% is the very good one. We classify the new distinct 382 test proteins into the α, β, α/β, α+β classes according to their SCOP classification. We checked whether our directional adjacency-score parameters could provide the success ratio of more than 90% for the different classes in the threading test. TABLE II shows such a success ratio for the proteins belonging to each class when they are subject to the threading test on the 382 test proteins.

Can the retro sequence of a protein in its native parent structure be stabilized?

A retro protein can be obtained by aligning a protein sequence backwards on its original native structure. The folding of the retro sequence of the B domain of Staphylococcal protein A was simulated by Kolinski and Skolnick (1994a, 1994b) and Olszewski et al. (1996) using a high coordination lattice model and the retro sequence was predicted to retain the structure close to the native parent structure or to a topological mirror image of it. However, Lacroix, Viguera, and Serrano (1998) showed the experimental evidence that the retro sequence of the SH3 domain (1SHG), the B domain of Staphylococcal protein A (1BDC), and the B1 domain of Staphylococcal protein G (2GB1) are unfolded proteins. As long as one ignores the backbone directionality in the coarse-graining description of a protein as a string of the isotropic beads, the protein score functions are insensitive to the sequence inversion and predict that the retro sequence fits well on the structure of its original native structures.

Our directional adjacency-score function was designed and built based on the backbone directionality of a protein, it is therefore expected to be sensitive to the sequence inversion of a protein. We applied our directional adjacency-score parameters to clarify whether the retro protein can retain its original structure as the stabilized one. Assuming that the retro sequence could be housed and stabilized in a native structure of its original sequence, we performed the threading test of the retro protein on our 1,006 training proteins in order to calculate the directional adjacency-score and check whether the retro protein could be stabilized. Table III shows the results of the threading test of the retro proteins of 1SHG, 1BDC, and 2GB1 using the directional adjacency-score parameters. The numbers of failed decoys for the original proteins with the forward sequences are 1, 0, and 0 which reflects that these are well stabilized in the conformational space of decoys. On the other hand the retro proteins with the backward sequences in a native structure of its original sequence are not stabilized at all with the large number of failed decoys whose directional adjacency-scores are lower than the score of the retro proteins. Therefore, our threading test using the directional adjacency-score function indicates at least that the retro sequences in their original native structures are not the stabilized one, which contradicts with the prediction of Kolinski and Skolnick (1994a; 1994b) and Olszewski et al. (1996) but agrees with the experimental evidence of Lacroix, Viguera, and Serrano (1998).

Conclusion and Prospects

We propose a new two-body score function which can stabilize all native states of 1,006 proteins simultaneously. This two-body score function differs from the usual pairwise contact function of Miyazawa and Jernigan (1985, 1996), Vendruscolo et al. (1999), Dima et al. (2000), and Salvi and DeLosRios (2003) in that it considers two amino acids which are adjacent each other at two ends of each peptide bond with the backbone directionality from the N-terminal to the C-terminal. The score is a corresponding propensity for such a directional alignment of two adjacent amino acids with their local environments. We show that the construction of a directional adjacency-score function was achieved using 1,006 training proteins with the sequence homology less than 30% which include all the representatives of different protein classes. After parameterizing the local environments of amino acids into 9 categories depending on three secondary structures and three

TABLE II. The success ratio for the proteins (from a set of 382 test proteins) belonging to α, β, α/β, α+β classes when they are subject to the threading test. It shows the success ratio of more than 90% for the different classes of proteins.

	α	β	α/β	A+β	total
The number of proteins	**87**	101	122	72	382
The number of failed proteins	5	4	0	2	11
Success Ratio (%)	94.3	96.0	100	97.2	97.1

kinds of hydrophobicity of amino acids, the 32,400 adjacency-score parameters of amino acids could be determined by the perceptron learning and protein threading. These could stabilize all the native folds of 1,006 training proteins simultaneously. When these parameters are tested on the new distinct 382 proteins with the sequence homology less than 90%, 371 (97.1%) proteins could recognize their native folds. Using these adjacency-score parameters, we showed that the retro sequence of the SH3 domain, the B domain of Staphylococcal protein A and the B1 domain of Streptococcal protein G can not be stabilized to fold, which agrees with the experimental evidence.

Materials and Methods

Perceptron learning

The general strategy to find the solution $\varepsilon_A \equiv \{\varepsilon_A(i,j;k,l)\}$ is to find the values of $\varepsilon_A(i,j;k,l)$ which satisfies Eq.(3) simultaneously for *D*=1,2,…,78.2 million in the 32,400-dimensional space of parameters;

$$\sum_{i,j=1}^{20}\sum_{k,l=1}^{9}[n_A(i,j;k,l)^D - n_A(i,j;k,l)]\varepsilon_A(i,j;k,l) = \vec{n}_A^D \cdot \vec{\varepsilon}_A > 0 \quad (3)$$

Here, $\vec{n}_A^D = [n_A(i,j;k,l)^D - n_A(i,j;k,l)]$ is fixed once a set of 1,006 training protein is known, and $\vec{\varepsilon}_A$ is the unknown vector to be determined. We start from an initial value of $\vec{\varepsilon}_A^0(i,j;k,l)$ and calculate the scalar product $\vec{n}_A^D$ on $\vec{\varepsilon}_A$ for all 78.2 million inequalities. The vectors $\vec{n}_A^D$ whose $\vec{n}_A^D \cdot \vec{\varepsilon}_A$ are negative are the ones which do not satisfy the above inequality and the corresponding decoys are called as the failed decoys. We select the worst vector $\vec{n}_A^w$ among the failed decoys, which has the lowest value of gap, and update $\vec{\varepsilon}_A(t+1) = \vec{\varepsilon}_A(t) + \alpha \cdot \vec{n}_A^w (0 < \alpha < 1)$ so that the gap for the worst decoy *w* increases. The 78.2 million scalar products are calculated again with the new $\vec{\varepsilon}_A(t+1)$, and the set of failed decoys and the worst decoy is identified to update $\vec{\varepsilon}_A(t+1)$ again. This procedure is iterated until the number of failed decoys out of 78.2 million decoys becomes zero. The main idea of this update is to find $\vec{\varepsilon}_A$ which can stabilize the score of the native states against the scores of the decoy structures so that the native state can be fully stabilized. If a solution of Eq.(3) exists, namely $\vec{\varepsilon}_A^{final}$ satisfies all 78.2 million inequalities, the vector $\vec{\varepsilon}_A^{final}$ converges to a region of points in the 32,400-dimensional space and the gap of the worst decoy $\vec{n}_A^w \cdot \vec{\varepsilon}_A^{final}$ becomes a positive finite within a finite number of iterations. If the iteration runs forever not to give a converging value of $\vec{\varepsilon}_A$ nor a positive finite value for the gap, the perceptron learning is not learnable which means that either there is no solution or the parameterization in the energy function (Eq.(1)) is not adequate.

TABLE III. The results of threading test for the retro proteins 1SHG, 1BDC, and 2GB1 using the directional adjacency-score function. It shows that the original proteins with the forward sequences are well stabilized, but the retro proteins with the backward sequences are not stabilized at all in the conformational space of decoys.

PDB	Total number of decoy	Number of failed decoy	
		Original protein	Retro protein
1SHG	195,362	1	37,998
1BDC	192,354	0	60,008
2GB1	196,366	0	1,132

Acknowledgement

This work is supported by the Creative Research Initiatives program (Center for Proteome Biophysics) of the ministry of education, science and technology/Korea Science and Engineering Foundation, Korea.

References

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science 181, 223-230.

Baker, D. (2000). A surprising simplicity to protein folding. Nature 405, 39-42.

Bowie, J.U., Luthy, R., Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164-170.

Bryant, S.H., Lawrence, C.E. (1993). An empirical energy function for threading protein sequence through folding motif. Proteins Struct. Funct. Genet. 16, 92-112.

Chang, I., Cieplak, M., Dima, R.I., Maritan, A., Banavar, J.R. (2001). Protein threading by learning. Proc. Natl. Acad. Sci. USA 98, 14350-14355.

Dima, R.I., Banavar, J.R., Maritan, A. (2000). Scoring functions in protein folding and design. Protein Sci. 9, 812-819.

Fersht, A.R. (1998). Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. Freeman, New York

Fischer, D., Rice, D.W., Bowie, J.U., Eisenberg, D. (1996) Assigning amino acid sequences to 3D protein folds. FASEB J. 10, 126-136.

Friedrichs, M.S., Wolynes, P.G. (1989). Toward Protein Tertiary Structure Recognition by means of Associative Memory Hamiltonians. Science 246, 371-373.

Godzik, A., Kolinski, A., Skolnick, J. (1992). Topology fingerprint approach to the inverse protein folding problem. J. Mol. Biol. 227, 227-238.

Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G. (1992). Protein tertiary structure recognition using optimized Hamiltonians with local interactions. Proc. Natl. Acad. Sci. USA 89, 9029-9033.

Heo, M., Cheon, M., Chang, I. (2004). Nonsymmetric two-body score function for protein fold recognition; Next nearest neighbor-adjacency of two amino acids. Int. J. Mod. Phys. C, 15, 1087-1094.

Heo, M., Kim, S., Moon, E.J., Cheon, M., Chung, K., Chang, I. (2005). Perceptron learning of pairwise contact energies for proteins incorporating the amino acid environment. Phys. Rev. E. 72, 011906/1-011906/9.

Hobohm, U., Sander, C. (1994). Enlarged representative set of

protein structures. Protein Sci. 3, 522-524.
Hobohm, U., Scharf, M., Schneider, R., Sander, C. (1992). Selection of representative protein data sets. Protein Sci. 1, 409-417.
Holm, L., Sander, C. (1996). Mapping the Protein Universe. Science 273, 595-602.
Jones, D.T., Taylor, W.R., Thonton, J.M. (1992). A new approach to protein fold recognition. Nature 358, 86-89.
Kolinski, A., Skolnick, J. (1994) a. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins Struct. Funct. Genet. 18, 338-352.
Kolinski, A., Skolnick, J. (1994)b. Monte Carlo simulation of protein folding. II. Application to protein A, ROP and crambin. Proteins Struct. Funct. Genet. 18, 353-366.
Krauth, W., Mezard, M. (1987). Learning algorithms with optimal stability in neural networks. J. Phys. A 20, L745-L752.
Lacroix, E., Viguerra, A.R., Serrano, L. (1998). Reading protein sequences backwards. Fold. Des. 3, 79-85.
Lee, B., Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. 55, 379-400.
Maiorov, V.N., Crippen, G.M. (1992). Contact potential that recognizes the correct folding of globular proteins. J. Mol. Biol. 227, 876-888.
Mayor, U., Guydosh, N.R., Johnson, C.M., Grossmann, J.G., Sato, S., Jas, G.S., Freund, S.M., Alonso, D.O., Daggett, V., Fersht, A.R. (2003). The complete folding pathway of a protein from nanoseconds to microseconds. Nature 421, 863-867.
Micheletti, C., Banavar, J.R., Maritan, A., Seno, F. (1998). Steric Constraints in Model Proteins. Phys. Rev. Lett. 80, 5683–5686.
Miyazawa, S., Jernigan, R.L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18, 534-552.
Olszewsk, K.A., Kolinski, A., Skolnick, J. (1996). Does a backwardly read protein sequence have a unique native state? Protein Eng. 9, 5-14.
Ouzounis, C., Sander, C., Scharf, M., Schneider, R. (1993). Prediction of protein structure by evaluation of equence-structure fitness. Aligning sequences to contact profiles derived from 3D structures. J. Mol. Biol. 232, 805-823.
Pattabiraman, N., Ward, K.B., Fleming, P.J. (1995). Occluded molecular surface: analysis of protein packing. J. Mol. Recognit. 8, 334-344.
Salvi, G., DeLosRios, P. (2003). Effective interactions cannot replace solvent effects in a lattice model of proteins. Phys. Rev. Lett. 91, 258102.
Seno, F., Vendruscolo, M., Maritan, A., Banavar, J.R. (1996). Optimal Protein Design Procedure. Phys. Rev. Lett. 77, 1901-1904.
Sippl, M.J., Weitckus, S. (1992). Detection of native like models for amino acid sequences of unknown three dimensional structure in a data base of known protein conformations. Proteins Struct. Funct. Genet. 13, 258-271.
Vendruscolo, M., Najmanovich, R., and Domany, E. (1999). Protein Folding in Contact Map Space. Phys. Rev. Lett. 82, 656–659.
Wilmanns, M., Eisenberg, D. (1993). Three-dimensional profiles from residue pair preferences: Identification of sequences with β/α-barrel fold. Proc. Natl. Acad. Sci. USA 90, 1379-1383.
Wolynes, P.G., Onuchic, J.N., Thirumalai, D. (1995). Navigating the folding routes. Science 267, 1619-1620.
Zhang, C., Kim, S.H. (2000). Environment-dependent residue contact energies for proteins. Proc. Natl. Acad. Sci. USA 97, 2550-2555.